A Collaborative Conversation Experimental Design for Real-Time Speech Adaptation Analysis

Jetic Gū[1,2], Samuel To[2], Ivan Fong[2], Fenqi Wang[2], Dawn Behne[4], Allard Jongman[5], Joan Sereno[5], Paul Tupper[3], and Yue Wang[2]

[1]School of Computing Science; [2]Department of Linguistics; [3]Department of Mathematics, Simon Fraser University;
[4]Department of Psychology, Norwegian University of Science and Technology; [5]Department of Linguistics, University of Kansas

1. Background and Design

Challenges

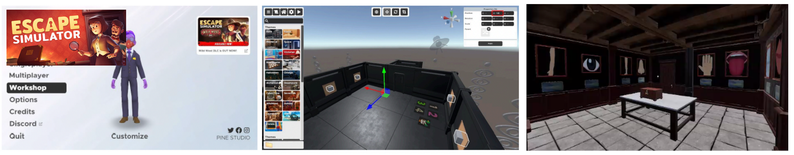
- **Phonetic Adaptation Experiments with Natural Conversations**
  - Designing an *engaging* task to elicit *natural/spontaneous* conversations[1]
  - Ensuring *repetition* of *target words* (sounds) without resorting to scripting[5]
  - *Synchronising* audio-video recording to capture both interlocutors
- **Human-Computer Interactions**
  - Generating *unscripted* yet controlled computer responses[3]
  - Recognising human speech and generating synthesised speech[6]

Our Design

- Employs a *collaborative video game* task to elicit natural conversations, where misproduction or misperception of target words may cause confusions, motivating adaptations
- Generates *unscripted real-time audio computer responses* for human-computer interactions during the game played online
- Enables studying *intelligibility-oriented phonetic adaptations* across speech contrasts and interlocutor backgrounds

2. Task Description

Task Creation in Escape Simulator



Escape Room game (3D)

- Two players must *collaboratively solve puzzles* in order to escape a "room" in which they are trapped. They
  - play in *separate rooms* in a video game;
  - *hear* each other and communicate verbally;
  - discuss *placement of pictures* (containing target words) on the wall;
  - may have to *make adjustments* in cases of *misunderstanding*.

Example

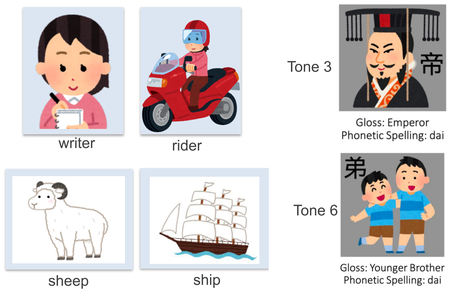
- Conversation between a vowel "merger" (Dawn=Don /ɑ/) and a "non-merger" (Dawn /ɔ/ vs. Don /ɑ/)

Non-merger:	What pictures are there on your floor?
Merger:	There is a picture of Don (/ɑ/), a hand, and a picture of Dawn (/ɑ/).
Non-merger:	Can you put the picture of Dawn (/ɔ/) on the top left corner?
Merger:	Are you talking about Don (/ɑ/) in blue shirt?
Non-merger:	I said Dawn (/ɔ/).

3. Studies

- **We have 3 ongoing studies using this experimental design**

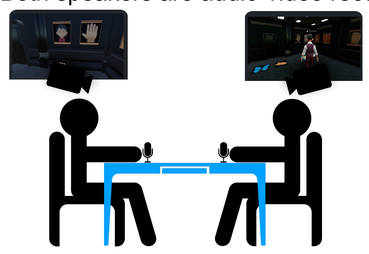
	Conversation dyad	Confusing target contrast
Study 1 (Wang et al., 5aSC)	Human vs Computer	Flap-stop (e.g., writer vs rider)
Study 2 (Zhang et al., 2pSCb)	Native (English) vs Nonnative (Mandarin)	English tense-lax vowels (e.g., sheep vs ship)
Study 3 (Fong et al., 4pSC)	Cantonese tone merger vs Non-merger	Cantonese Tone 3 & Tone 6 (e.g. dai3 "emperor" vs dai6 "brother")

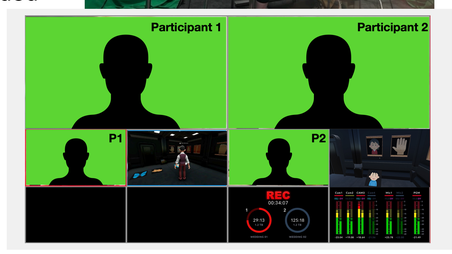


4. Recording Setup

Human-Human Conversation

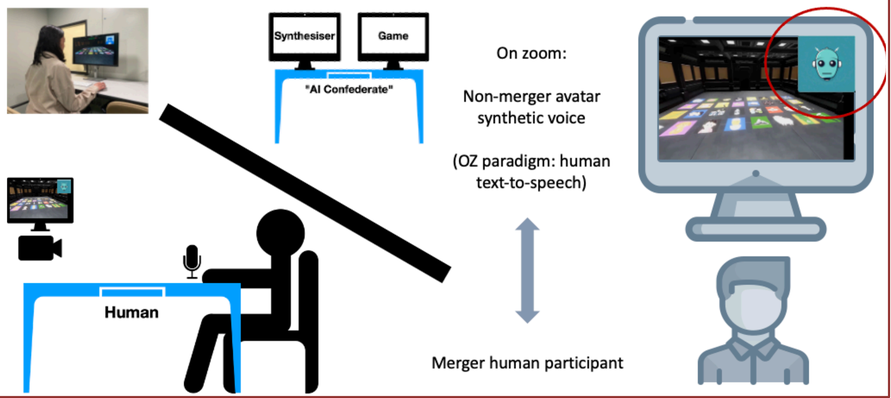
- Two speakers of different backgrounds playing Escape room game (e.g., vowel merger vs non-merger)
- Speakers face one another, each in front of a microphone and keyboard, with monitor and camera mounted behind them
- Both speakers are audio-video recorded





Human-Computer Conversation

- **Wizard-of-Oz** paradigm
  - **Participant**: *human speaker* playing Escape room game on Zoom with AI
  - **Confederate**:
    - "AI" controlled by *human confederate* who types responses behind-the-scene (programmed to autocomplete sentences to speed up input)
    - *Speech synthesiser*[4] converts text to speech, providing real-time computer responses (programmed to simulate a desired accent)



5. Summary and Future Directions

Challenges in Conversational Phonetic Adaptation

1. Our experimental design is *unscripted*, while ensuring *repeated utterances* of target words for acoustic analysis
2. Since the entire conversation is recorded, we are able to observe the *dynamicity of speech adjustments* over time
3. The current task is designed for studying *phonetic adaptations* for *intelligibility* benefit
4. We conducted experiments in *Human-Human* and *Human-AI* interactions. See our other posters (Studies 1-3) for more detail

Future Directions

1. Evaluate the current design not only through acoustic analysis but also through *intelligibility testing*
2. Create *adaptive AI-powered vocal interface* that can make adjustments when human speakers misunderstand
3. Extend the current analysis which focuses on phoneme-specific adaptations to *broader-domain conversational analysis* (e.g., the role of global prosodic and durational features in adaptations for intelligibility gain).

The current design that elicits natural conversations enables analysis of language-specific and language-universal phonetic adaptations across human and AI interlocutors.

References

1. Kasper, G., & Wagner, J. (2014). Conversation Analysis in Applied Linguistics. *Annual Review of Applied Linguistics*, 34, 171–212
2. Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. *In Studies in the organization of conversational interaction* (pp. 7-55). Academic Press.
3. Bell, L., Gustafson, J., & Heldner, M. (2003, August). Prosodic adaptation in human-computer interaction. *In Proceedings of ICPHS* (Vol. 3, pp. 833-836). Citeseer.
4. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. *In International Conference on Machine Learning* (pp. 28492-28518). PMLR.
5. Lewandowski, N., & Jilka, M. (2019). Phonetic convergence, language talent, personality and attention. *Frontiers in Communication*, 4, 18.
6. Bell, L. (2003). *Linguistic Adaptations in Spoken Human-Computer Dialogues-Empirical Studies of User Behavior* (Doctoral dissertation, Institutionen för talöverföring och musikakustik).
7. Biro, T., Toscano, J. C., & Viswanathan, N. (2022). The influence of task engagement on phonetic convergence. *Speech Communication*, 138, 50-66.
8. Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., ... & Yu, Z. (2022). Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, 71, 101255.

Acknowledgements

We thank members of the Language and Brain Lab and the Tech Team at Simon Fraser University for their assistance. This project is funded by NSERC Discovery Grant 2023-05666 to Y. Wang.